CS4780 - Machine Learning

Fall 2009

Thorsten Joachims
Cornell University

Outline of Today

- **Who we are?**
- **What is learning?**
- **Why should a computer be able to learn?**
- **Examples of machine learning**
- **What it takes to build a learning system?**
- **Syllabus**
- **Administrivia**
  - Pre-Requisites
  - Assignments
  - Grading
  - Textbook and course material
  - Office Hours

(One) Definition of Learning

Definition [Mitchell]:
A computer program is said to learn from
- experience E with respect to some class of
- tasks T and
- performance measure P,

if its performance at tasks in T, as measured by P, improves with experience E.

Syllabus

- **Concept Learning :** Hypothesis space, version space, target concept
- **Instance-Based Learning :** K-nearest neighbor, collaborative filtering
- **Decision Trees :** TDIDT, Representation bias vs. search bias
- **Hypothesis Tests :** Confidence intervals, resampling estimates
- **Linear Rules :** Perceptron, Winnow
- **Support Vector Machines :** Optimal hyperplane, Kernels
- **Generative Models :** Naïve Bayes, MAP and Bayesian learning
- **Hidden Markov Models :** Viterbi, Expectation-Maximization
- **Complex Output Prediction :** natural language parsing
- **Learning Theory :** PAC learning, Mistake Bounds, No-Free-Lunch
- **Clustering :** HAC, k-means, latent semantic indexing

Textbook and Course Material

- **Main Textbook**
  - Main: Tom Mitchell, "Machine Learning", McGraw Hill, 1997.
  - Cristianini, Shawe-Taylor, "Introduction to Support Vector Machines", Cambridge University Press, 2000. (online)
  - Schoelkopf, Smola, "Learning with Kernels", MIT Press, 2001. (online)
  - Course pack (one chapter)
- **Additional Reference (optional)**
  - Ethem Alpaydin, "Introduction to Machine Learning", MIT Press, 2004.
- **Course Notes**
  - Slides available on course homepage
  - Material on blackboard

Pre-Requisites and Related Courses

- **Pre-Requisites**
  - Programming skills (e.g. CS 2110)
  - Basic linear algebra (e.g. MATH2940)
  - Basic probability theory (e.g. CS 2800).
- **Related Courses**
  - CS4700: Foundations of Artificial Intelligence
  - CS4300: Information Retrieval
  - CS6780: Advanced Machine Learning
  - CS678?: Advanced Topics in Machine Learning
  - CS6740: Advanced Language Technologies

Assignments and Grading

- **Deliverables**
  - 2 Prelim Exams (40% of Grade)
  - Final Project (15% of Grade)
  - Homeworks (~5 assignments) (40% of Grade)
  - Class Participation (5% of Grade)
- **Policies**
  - Assignments are due at the beginning of class on the due date.
  - Assignments turned in late will drop 5 points for each period of 24 hours for which the assignment is late.
  - No assignments will be accepted after the solutions have been made available.
  - Collaborations are not allowed (except when explicitly permitted).
  - Must state all sources of material used in assignments or project.
  - Academic Integrity

Final Project

- **Organization**
  - Self-defined topic related to your interests and research
  - Groups of 3-4 students
  - Each group has TA as advisor
- **Deliverables**
  - Project proposal (~ week after spring break)
  - Meetings with TA to discuss progress
  - Short presentation in class (last week of classes)
  - Project report (~ exam period)

How to Get in Touch

- **WWW Page**
  - http://www.cs.cornell.edu/Courses/cs4780/2009fa/
- **Email Addresses**
  - Thorsten Joachims: tj@cs.cornell.edu
  - Mark Verheggen: mark@cs.cornell.edu
  - Rick Ducott, Haden Hooyeon Lee, Vaibhav Goel
  - Mailing list to all course staff: cs4780-l@lists.cs.cornell.edu (?)
- **Office Hours**
  - Thorsten Joachims:
    - Tuesdays 4:30pm – 5:30pm, 4153 Upson Hall (not 9/1)
  - Other office hours:
    - TBD